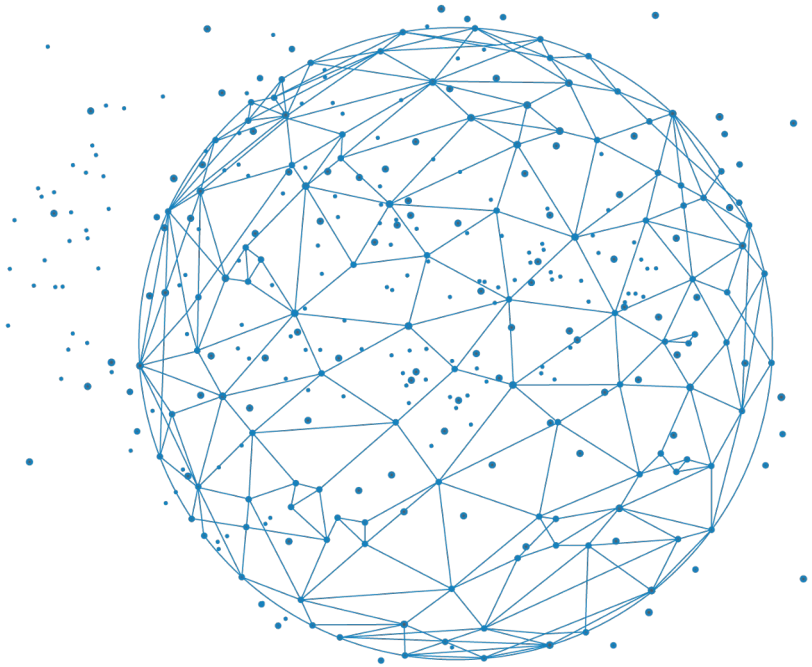


Introduction à la reproductibilité en bioinformatique

V. Cognat, IBMP

Slides from IFB - FAIR bioinfo

BiGEst

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

cnrs INRAE Inserm cea

Plan

Partie 1 : Contexte

- Science ouverte
- Crise de la reproductibilité

Partie 2 : Les dispositifs nationaux

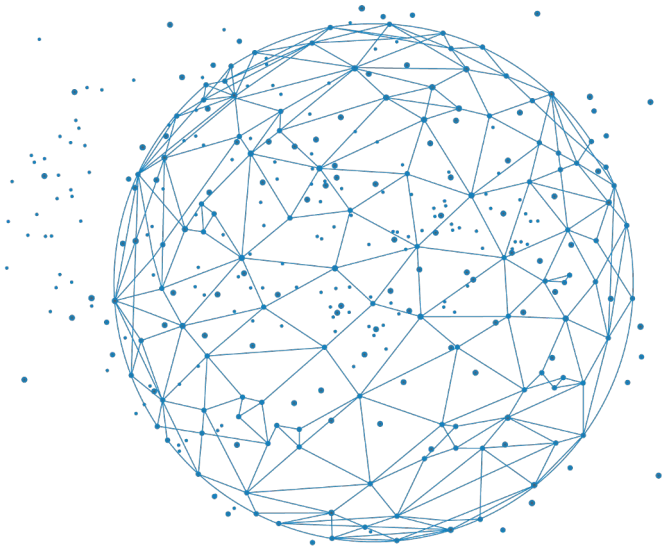
- Les 2 plans nationaux
- Des comités et ressources nationales

Partie 3 : Les concepts de base

- Les définitions de la reproductibilité
- Les principes FAIR appliqués aux logiciels de la recherche

Partie 4 : Une proposition de solution

Contexte
Science Ouverte, crise de la reproductibilité



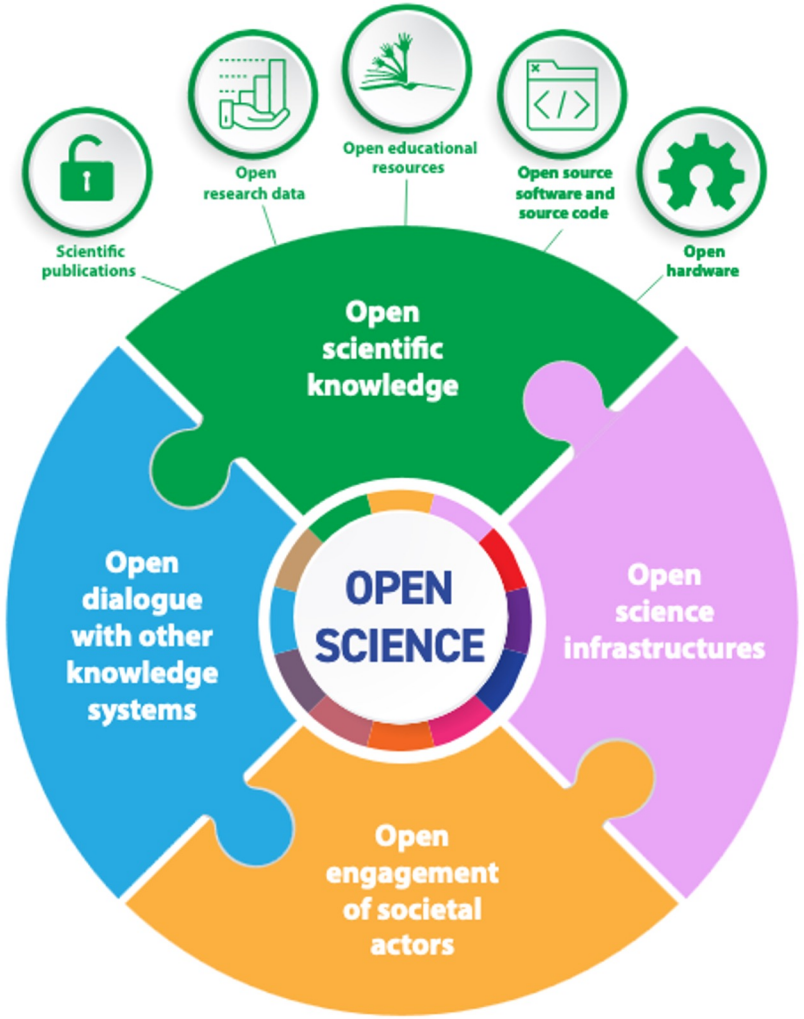
Contexte: Science Ouverte

Plusieurs facteurs :

- **Révolution numérique** : capacité de partager la connaissance
- Pression sur l'**innovation** et la **collaboration** : nombreux défis mondiaux qui nécessitent des approches multidisciplinaires et collaboratives (changement climatique, santé humaine, …)
- Des demandes croissantes de **transparence** et de redevabilité par le grand public (scandales scientifiques)
- Coûts de publications et de l'**accès aux revues scientifiques**
- Maximiser l'**impact** de la recherche (domaines d'applications, grand public)

Contexte: Science Ouverte

Recommandations de l'UNESCO



UNESCO Recommendation on Open Science, nov 2021
https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

Objectif : rendre la recherche accessible à tous

- Pas seulement l'accès à la **connaissance** elle-même
- Tout le processus de **création** et de **dissémination** de celle-ci
- La possibilité de **réutilisation**
- Ouverture au **dialogue** avec tous les acteurs, interdisciplinarité
- Engagement de et vers la **société**

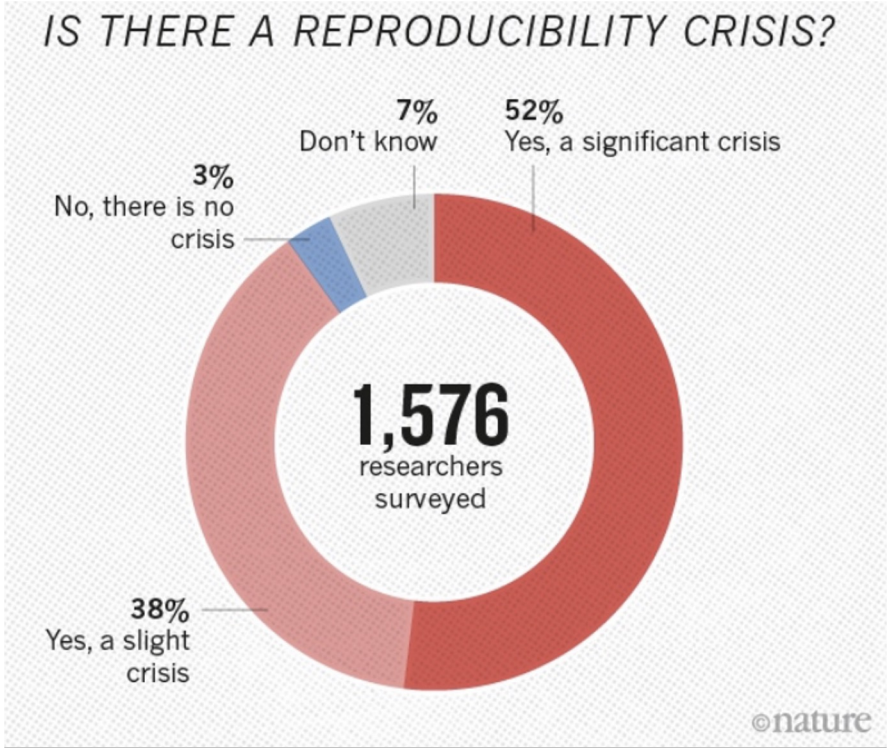
Ambitions

- **Démocratiser l'accès aux savoirs**
- **Rendre la science plus cumulative**, plus fortement étayée par les données, plus transparente
- **Augmenter l'efficacité de la recherche** en évitant de dupliquer les efforts, en ré-utilisant des données ou du matériel scientifique
- **Favoriser les avancées scientifiques** et l'**innovation**
- Favoriser la **confiance des citoyens dans la science**

La crise de la reproductibilité

90%

des chercheurs reconnaissent une crise de la reproductibilité en science



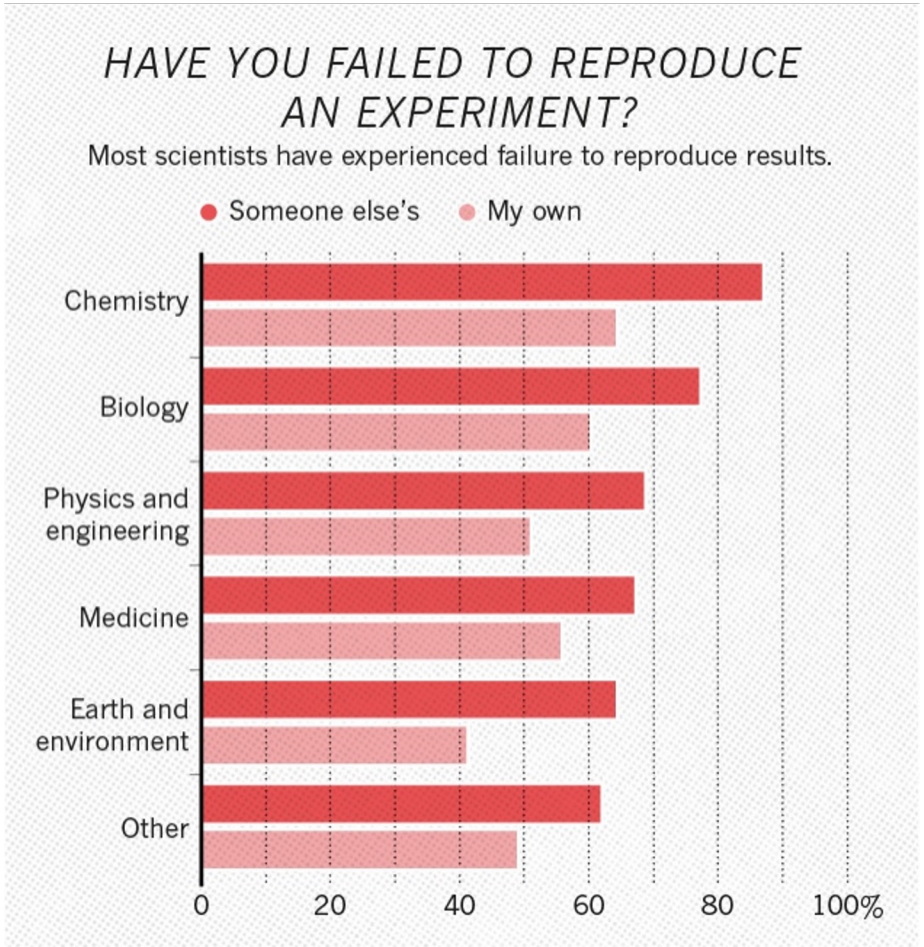
Monya Baker, Nature 2016
https://doi.org/10.1038/533452a

La crise de la reproductibilité : en biologie

En biologie

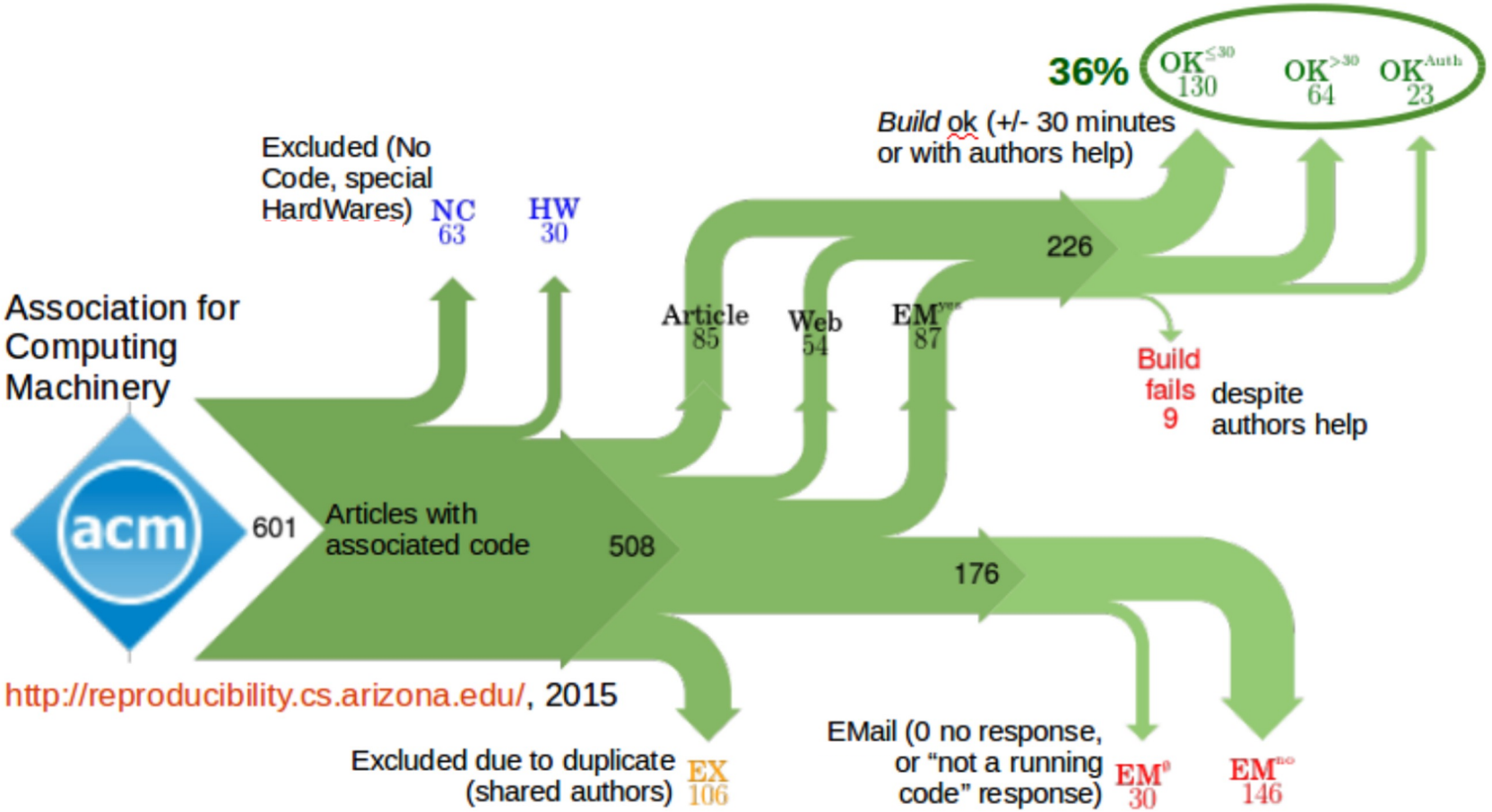
76 %

des chercheurs interrogés ont échoué à reproduire des résultats



Monya Baker, 2016
https://doi.org/10.1038/533452a

La crise de la reproductibilité : en informatique



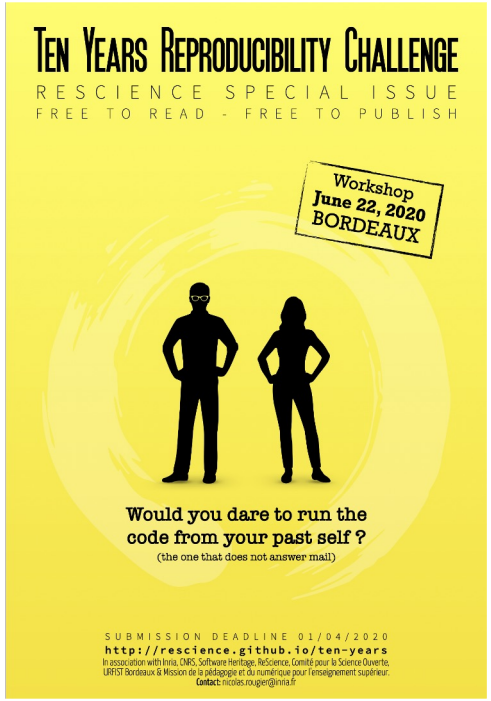
64% failure to reuse the software

(Collberg et al. 2015)

Un problème de reproductibilité aussi en bioinformatique

Qui n'a jamais voulu ré-utiliser un protocole, un pipeline, un outil ou un jeu de données sans y arriver ??

- Documentation obsolète ou URL not found (E404!!!)
- Outils : OS non compatible, dépendances pas disponible, mise à jour ⇒ codes inutilisables : python 2 vs. 3, changement d'arguments pour des fonctions (R)
- Impossibilité de reproduire les résultats de l'analyse computationnelle : versions de paquets, IDE : version stable de la langue différente selon le système d'exploitation (Rstudio)

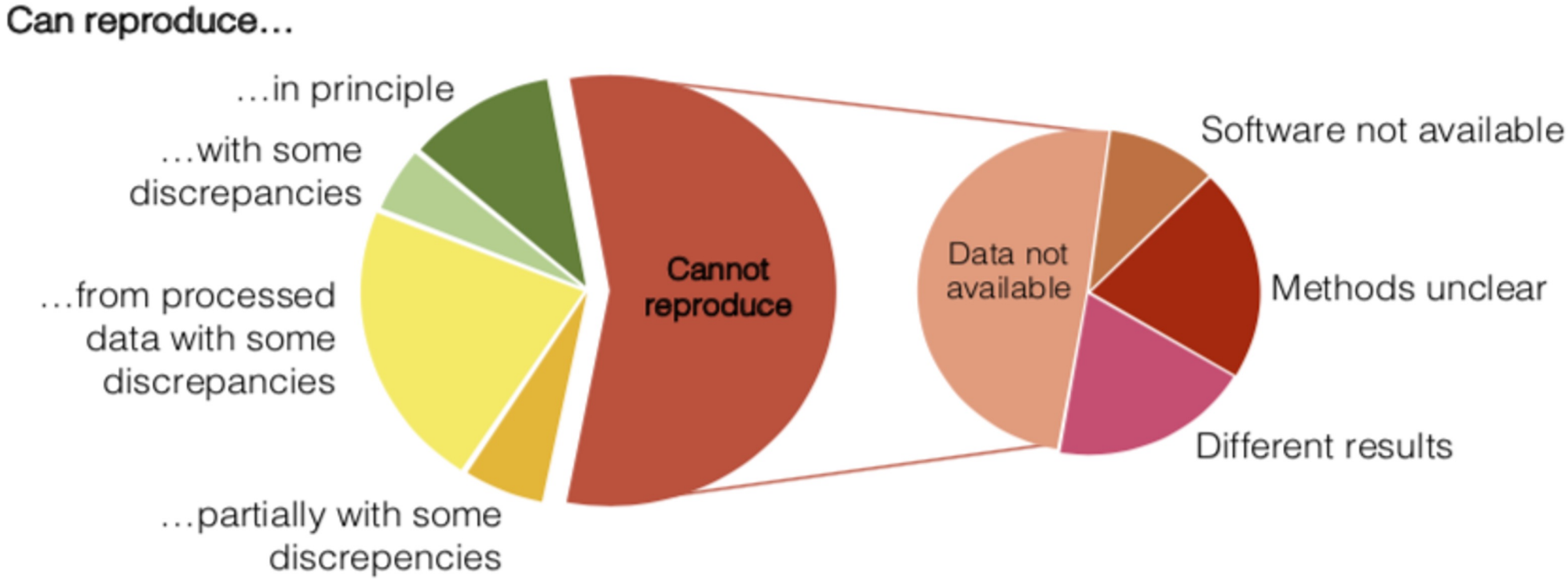


https://rescience.github.io/

Ten-Year Reproducibility Challenge, Konrad Hinsen Can your 2009 code still run? special issue of ReScience and result comments in Nature

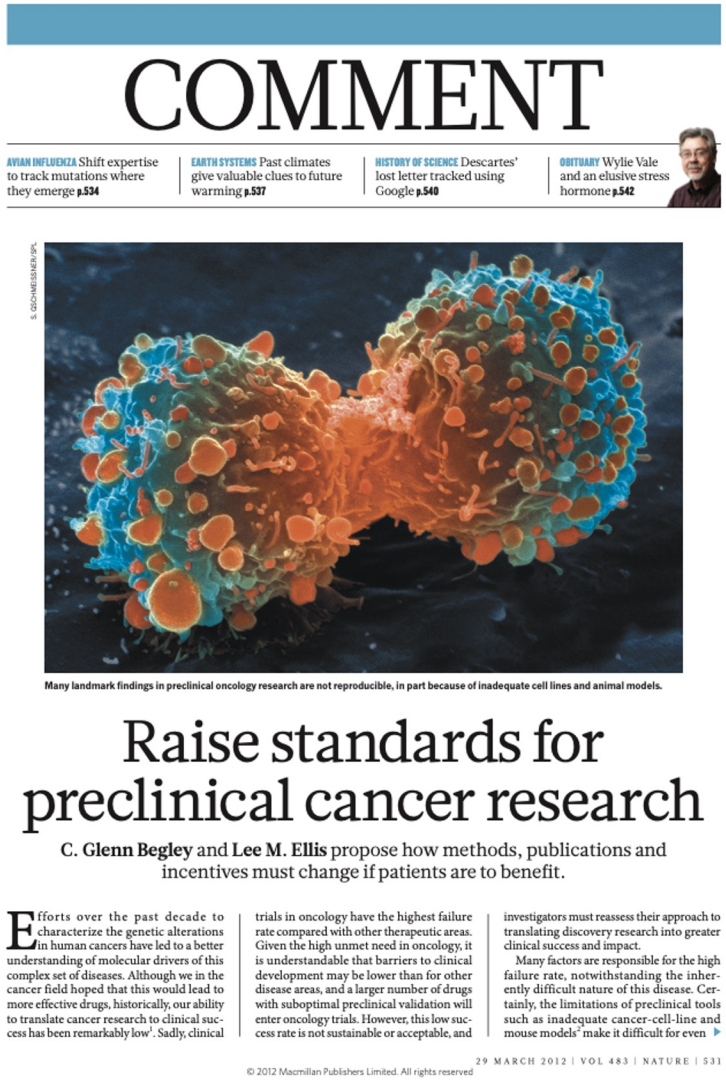
La crise de la reproductibilité : en bioinformatique

Replication of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006:



Summary of the efforts to replicate the published analyses.
Adopted from: Ioannidis et al. Repeatability of published microarray gene expression analyses.
Nature Genetics **41** (2009) doi:10.1038/ng.295

La crise de la reproductibilité : en bioinformatique



COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534 | EARTH SYSTEMS Past climates give valuable clues to future warming p.537 | HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540 | OBITUARY Wylie Vale and an elusive stress hormone p.542

S. GSCHMEISSNER/SPL

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

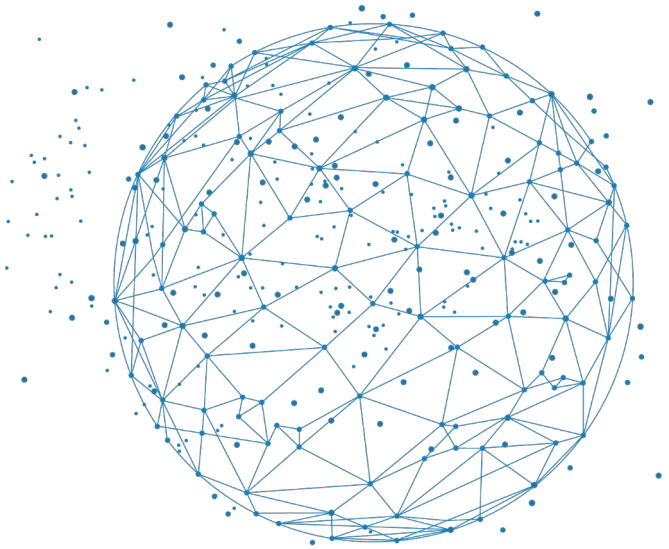
Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models[2] make it difficult for even ▶

29 MARCH 2012 | VOL 483 | NATURE | 531

Alsheikh-Ali et al. PLoS ONE (2011)
Nekrutenko & Taylor, Nature Genetics (2012)
Begley & Ellis Nature (2012)
…

Les dispositifs nationaux



Les données de la recherche sont des informations publiques :

- Principe d'ouverture par défaut et de libre utilisation (Loi Lemaire - Loi Pour une République Numérique 2016)
- Principe de gratuité (Loi Valter 2015) :
  - Seule une liste fermée d'administrations peuvent fixer des redevances de réutilisation (IGN, Météo France)
  - Articulation possible avec le dépôt de brevets et d'autres formes de valorisation

« aussi ouvert que possible,

aussi fermé que nécessaire »

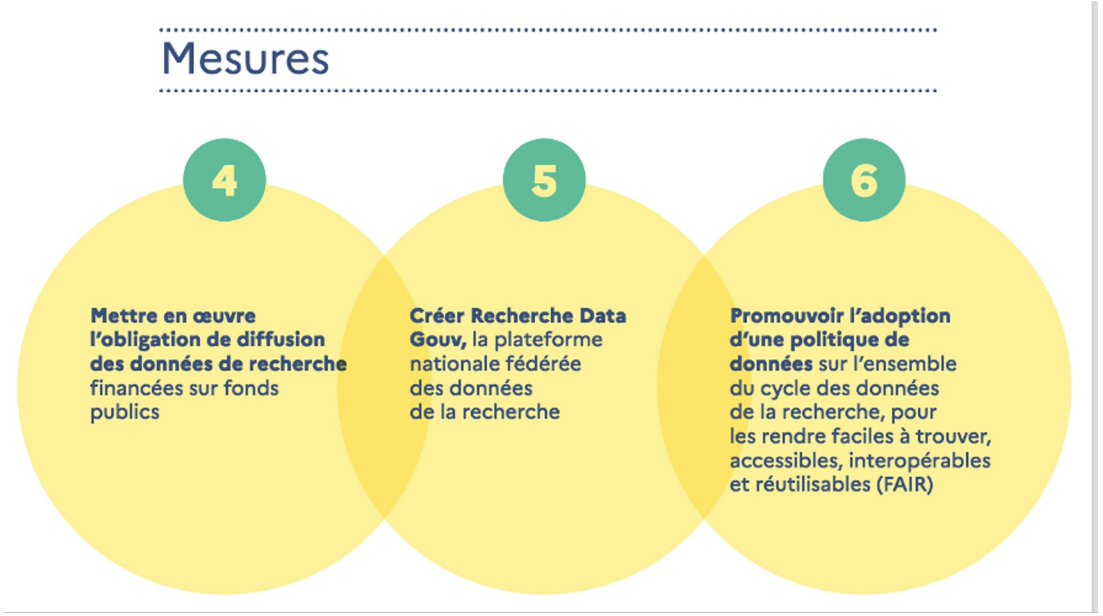
Deux plans nationaux pour la Science Ouverte

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Premier plan national pour la science ouverte : poser les principes de la science ouverte

2018-2021

2021-2024



Deuxième plan national pour la science ouverte : périmètre étendu aux algorithmes et codes sources et déclinaisons thématiques

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE
elixir FRANCE
BiGEst

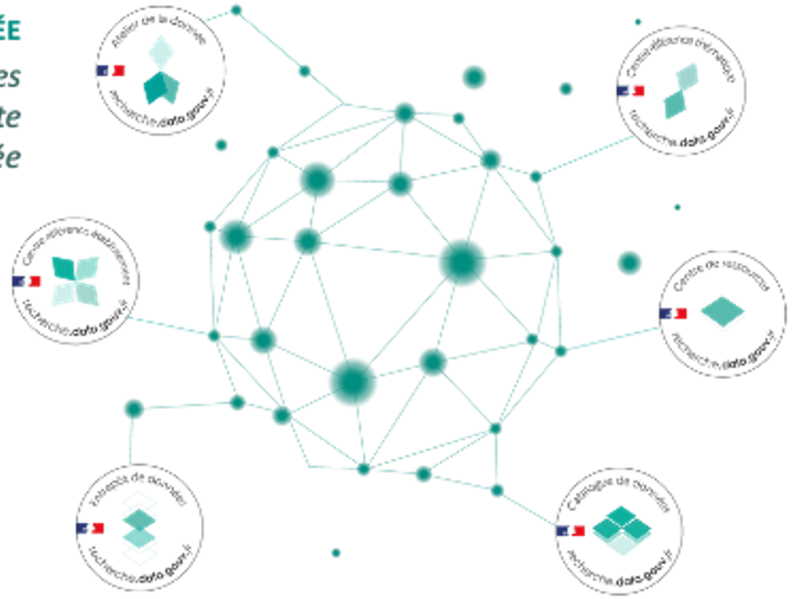
MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Un écosystème au service du partage et de l'ouverture des données de la recherche

15 + ATELIERS DE LA DONNÉE
Expertise généraliste en proximité des équipes de recherche pour toute question relative à la donnée

4 CENTRES DE RÉFÉRENCE ÉTABLISSEMENT
Accompagnement propre aux orientations de l'établissement

1 ENTREPÔT DE DONNÉES
Offre mutualisée pour tous les établissements pour le dépôt et la publication des données



6 + CENTRES DE RÉFÉRENCE THÉMATIQUES
Expertise par domaine scientifique

4 CENTRES DE RESSOURCES
Pour soutenir les ateliers et capitaliser leurs pratiques

1 CATALOGUE DES DONNÉES
Repérer et signaler les données des entrepôts externes de confiance

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE
elixir FRANCE
BiGEst

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Les ateliers de la donnée

décembre 2022

Expertise généraliste en proximité des équipes de recherche pour toute question relative à la donnée

Mutualisation des services et des compétences des établissements à l'échelle d'un territoire
Association des métiers de recherche / BU-IST / informatique

Développement progressif des ateliers de la donnée
- *Au rythme de leur conception par les établissements*

adele HELPDESK

Les centres de référence thématiques

novembre 2022

Expertise par domaine scientifique

Définition et diffusion des **bonnes pratiques et des standards internationaux de gestion, traitement et diffusion des données**

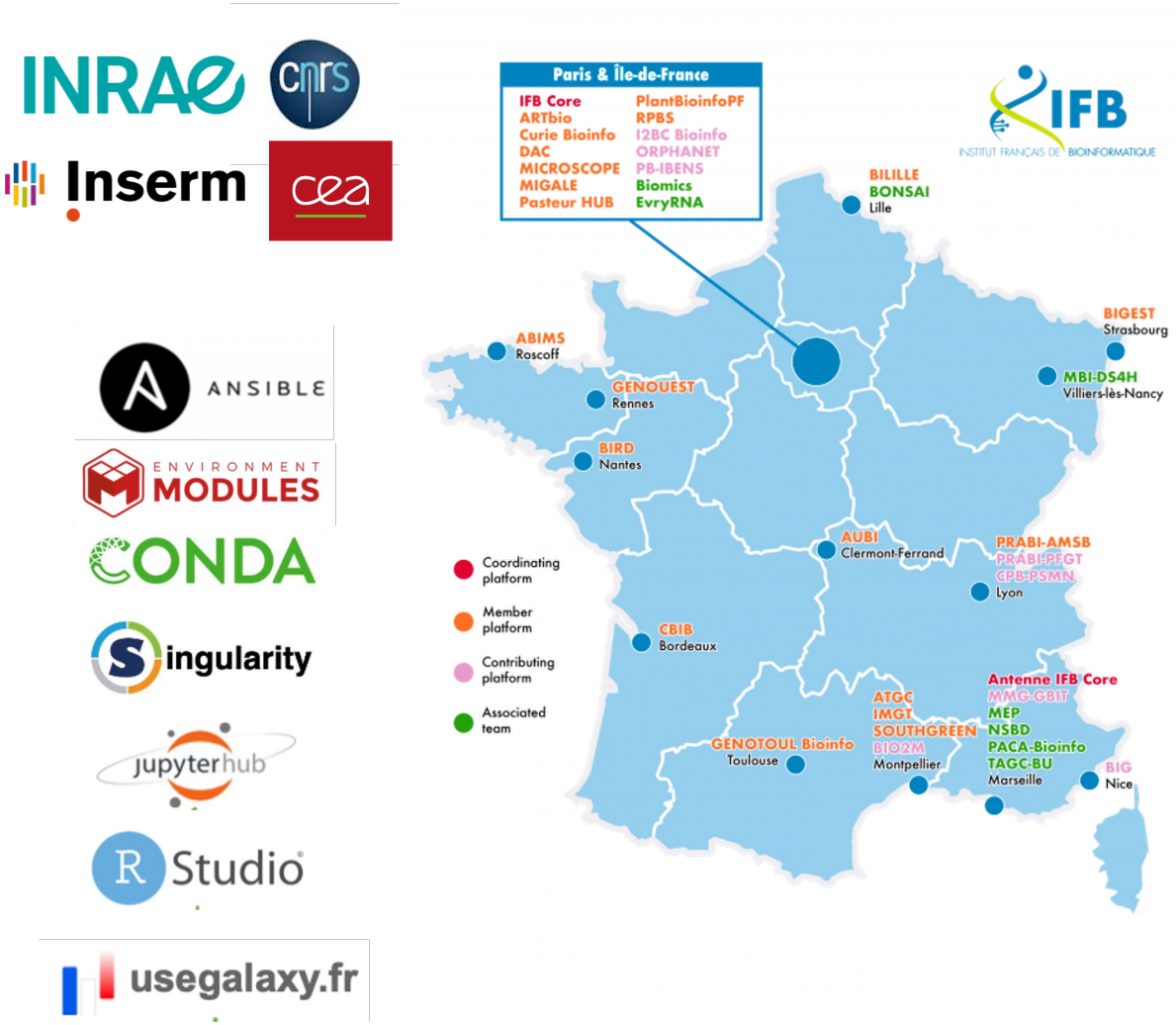
par domaine scientifique

Premier paysage proposé :

infrastructures de recherche ayant une activité structurante de gestion et diffusion de données pour leur communauté scientifique

L’IFB est une infrastructure numérique distribuée multi-tutelles qui propose

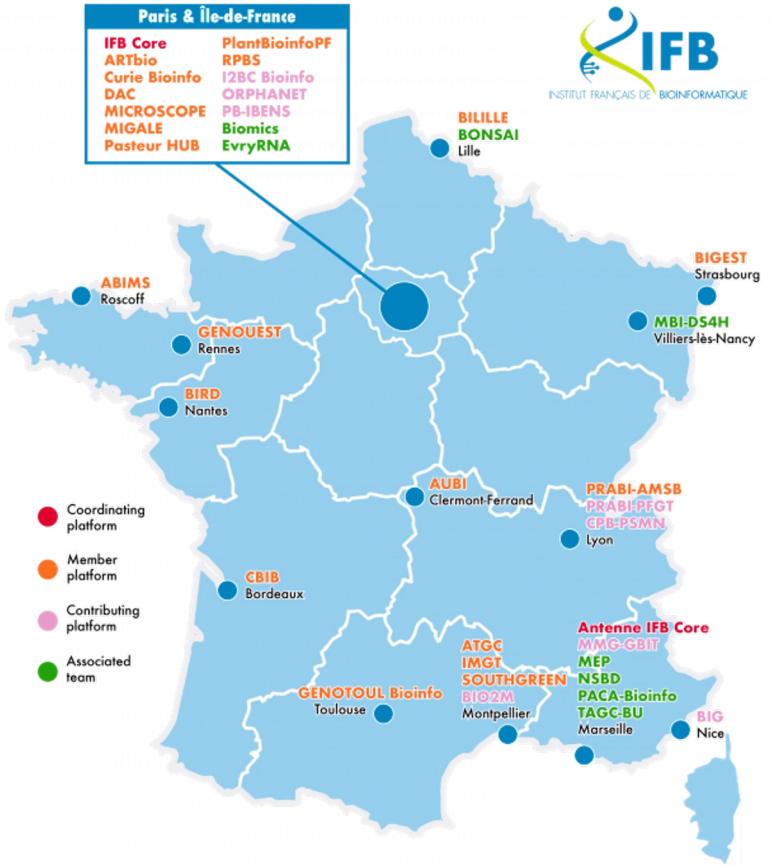
- des **clusters et clouds nationaux** avec des environnements de travail facilitant la **reproductibilité**
- des **outils** et **banques de données**
- **des formations**



L'IFB : Centre de Ressource Thématique en Biologie Santé

En tant que Centre de Ressource thématique

- Une aide à la définition et à la diffusion de **standards** et **de référentiels** en biologie - santé
- Des développements logiciels dédiés à la **science ouverte,** ex: **FAIRchecker**
- La conception et diffusion de **tutoriels** et formations aux **principes FAIR et à la reproductibilité**



FAIR-Checker
https://fair-checker.france-bioinformatique.fr

Accompagnement

- **Les Référents Données dans les Instituts (RDS et RDO) ou data steward**
- **Le délégué à la protection des données (DIL) ou Data Protection Officer (DPO)**
- **Les Administrateurs des Données Algorithmes et Codes de la recherche (ADAC)**
- **Les Responsables de la Sécurité des Systèmes d'Information (RSSI)**

Et aussi: responsables partenariats et valorisation, responsables qualité,..

Bonnes pratiques de gestion

Protection

Gouvernance

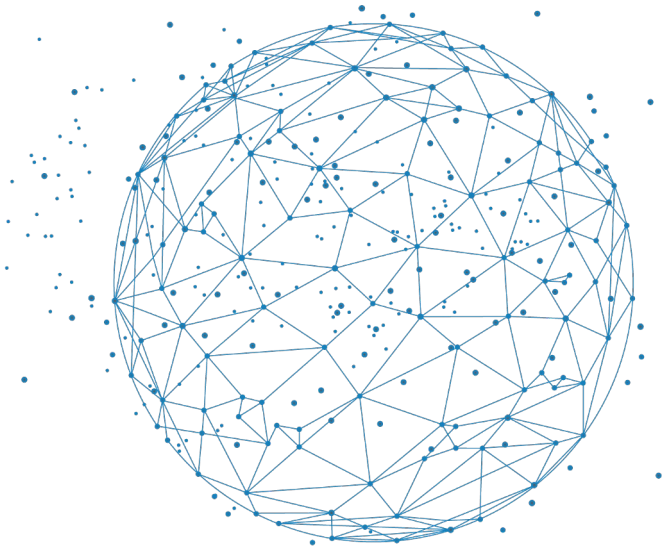
Sécurité

Valeur, Qualité

Des ressources utiles pour aller plus loin

- Recherche Data Gouv
  - Classes virtuelles et tutoriels
- DoRANum
  - Les fiches techniques
- Ouvrir la Science
  - dont les passeports pour la Science Ouverte

Les concepts de base de la reproductibilité



La reproductibilité en sciences

Reproducible research, Repeatability, Replicability, Reproducibility, …

concepts chevauchants ⇒ Attention aux définitions!

https://www.researchgate.net/publication/323118701_Terminologies_for_Reproducible_Research

Définitions de Association for Computing Machinery (2016) :

Repeatability même équipe, même design expérimental

Reproducibility équipes différentes, même design expérimental

Replicability équipes différentes, designs différents

https://www.acm.org/publications/policies/artifact-review-and-badging-current

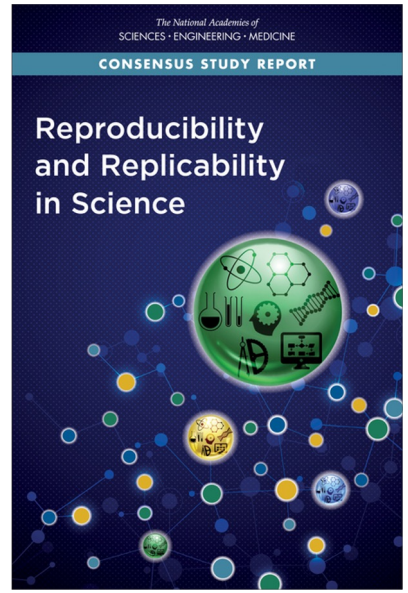
Matrice pour la reproductibilité de Whitaker (2017):

		Data: Same	Data: Different
Code: Same		Reproducible	Replicable
Code: Different		Robust	Generalisable

https://doi.org/10.6084/m9.figshare.5443201.v1, Slide number 7

National Academies of Sciences, Engineering, and Medicine (2019) :

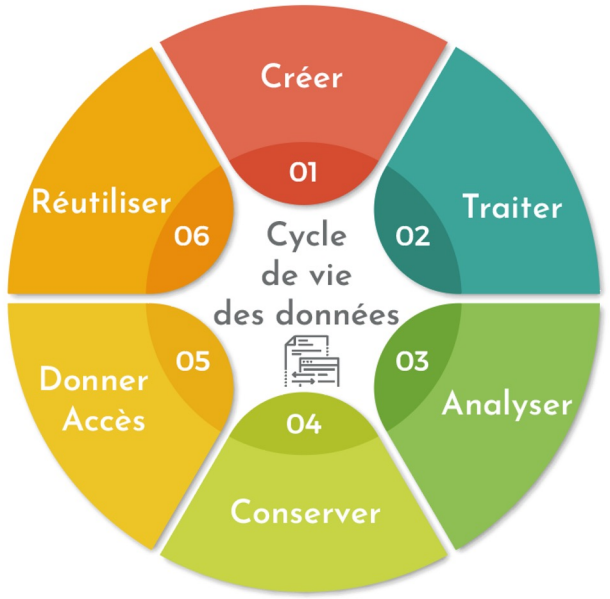
Washington DC.
The National Academies Press,
https://www.nap.edu/read/25303/chapter/1



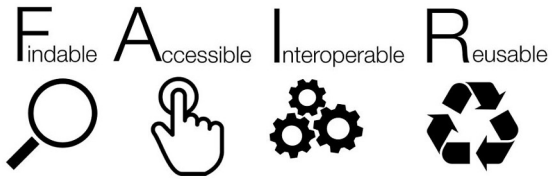
Data & Code

Code	Data: Same	Data: Different
Same	Reproducible	Replicable
Different	Robust	Generalisable

?



https://science-ouverte.univ-artois.fr/les-donnees-de-la-recherche/



Code/Logiciel: une donnée spéciale

Comme les données, les logiciels sont sauvegardés mais ils ne sont pas “juste” des données

Ce sont des objets **vivants** et **complexes**

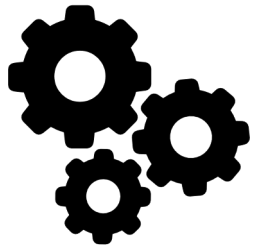
	Données	Logiciels
sont	des faits, des observations	des créations
produisent	des preuves	un outil exécutable
changent ?	non (sauf modif. du hardware sous-jacent)	oui : maintenance & mise à jour continues
durée de vie	au long terme nécessaire quand l’expérience ne peut être reproduite (coût de collecte ou de validation)	courte : souvent reconstruits pour utiliser d’autres logiciels (dépendances complexes), remplaçables (un autre réalise mieux la tâche)

Katz DS, Niemeyer KE, Smith AM, Anderson WL, Boettiger C, Hinsen K, Hooft R, Hucka M, Lee A, Löffler F, Pollard T, Rios F. 2016. Software vs. data in the context of citation. PeerJ Preprints 4:e2630v1 https://doi.org/10.7287/peerj.preprints.2630v1

Les principes FAIR

Findable **A**ccessible **I**nteroperable **R**eusable






FAIR for research software

FAIR principles[1] as listed by GO FAIR	FAIR research software principles,[4] changes are underlined
F. Findable	**F. Findable**
The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.	The first step in (re)using software is to find it. Metadata and software should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of software, so this is an essential component of the FAIRification process.
F1. (Meta)data are assigned a globally unique and persistent identifier	F1. Software is assigned a globally unique and persistent identifier
F2. Data are described with rich metadata (defined by R1 below)	F2. Software is described with rich metadata (defined first by R1 below, and then by the original FAIR principles for metadata)
F3. Metadata clearly and explicitly include the identifier of the data they describe	F3. Metadata clearly and explicitly include the identifier of the software they describe
F4. (Meta)data are registered or indexed in a searchable resource	F4. Software is registered or indexed in a searchable resource
A. Accessible	**A. Accessibl**
Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.	Once the use accessed, possibly including authentication and authorization.
A1. (Meta)data are retrievable by their identifier using a standardized communications protocol	A1. Software protocol
A1.1. The protocol is open, free, and universally implementable	A1.1. The pr
A1.2. The protocol allows for an authentication and authorization procedure, where necessary	A1.2. The pr where neces
A2. Metadata are accessible, even when the data are no longer available	A2. Metadat
I. Interoperable	**I. Interoperabl**
The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.	The software us metadata). Soft
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	(deemed unnec
I2. (Meta)data use vocabularies that follow FAIR principles	(deemed unnec
R1.3. [(Meta)data meet domain-relevant community standards] used as a model for a new Interoperability guiding principle	I1. Software sh standards
I3. (Meta)data include qualified references to other (meta)data	I2. Software inc
R. Reusable	**R. Reusable**
The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.	The ultimate go this, software sh replicated, com
R1. (Meta)data are richly described with a plurality of accurate and relevant attributes	R1. Software is richly described with a plurality of accurate and relevant attributes
R1.1. (Meta)data are released with a clear and accessible data usage license	R1.1. Software is made available with a clear and accessible software usage license
R1.2. (Meta)data are associated with detailed provenance	R1.2. Software is associated with detailed provenance
R1.3. (Meta)data meet domain-relevant community standards	R1.3. Software meets domain-relevant community standards
I3. [(Meta)data include qualified references to other (meta)data] used as a model for a new Reusability guiding principle	R2. Software includes qualified references to other software

Opinion

Taking a fresh look at FAIR for research software

Daniel S. Katz,[1,*] Morane Gruenpeter,[2] and Tom Honeyman[3]
[1]University of Illinois, Urbana, IL, USA
[2]Inria, Software Heritage, Paris, France
[3]Australian Research Data Commons, Melbourne, VIC 3145, Australia
*Correspondence: d.katz@ieee.org
https://doi.org/10.1016/j.patter.2021.100222

FAIR principles[1] as listed by GO FAIR	FAIR research software principles,[4] changes are underlined
The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.	The first step in (re)using software is to find it. Metadata and software should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of software, so this is an essential component of the FAIRification process.
F1. (Meta)data are assigned a globally unique and persistent identifier	F1. Software is assigned a globally unique and persistent identifier
F2. Data are described with rich metadata (defined by R1 below)	F2. Software is described with rich metadata (defined first by R1 below, and then by the original FAIR principles for metadata)
F3. Metadata clearly and explicitly include the identifier of the data they describe	F3. Metadata clearly and explicitly include the identifier of the software they describe
F4. (Meta)data are registered or indexed in a searchable resource	F4. Software is registered or indexed in a searchable resource

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7961177/

Et dans la vraie vie ?

Facile à trouver (pour l'Homme et la machine)

- Identifiant unique (un DOI par exemple avec **Zenodo**)
- Métadonnées décrivant l'analyse et les outils ; elles sont FAIR, consultables et indexables (exemple **README**)
- Disponible sur une forge (**GitHub**, **GitLab**) et sur un entrepôt d'archives (**SWH**)

Accessible

- **Licence** et droit d'accès (selon employeur, DPO)
- Métadonnées accessibles, même lorsque le logiciel n'est plus disponible (**SWH**)

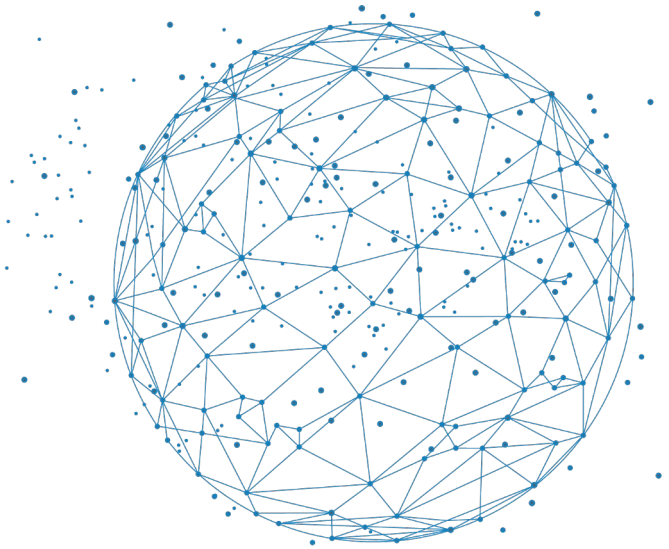
Interopérable

- Coopération des outils (**snakemake**, **conda** et **docker**) en local ou serveur (cloud ou cluster)

Réutilisable

- Protocole rejouable (**snakemake**) à l'identique (**Jupyter**, **Quarto**) dans un environnement virtuel (**Conda**, **Docker, Apptainer**)

Une proposition de solution pour rendre reproductible un projet de bioinformatique



Les piliers de la reproductibilité

Reproductibilité

Versionner Partager Archiver	Créer un environnement virtuel	Installer des outils par un gestionnaire	Créer un script d'analyses	Déporter l'analyse sur un serveur	Portabilité de l'exploration des résultats	Éditer un rapport d'analyses

d'après
The five pillars of computational reproducibility: Bioinformatics and beyond
Mark Ziemann, Pierre Poulain, Anusuiya Bora, OSF preprint, 2023

Et vous ? Combien de piliers soutiennent votre reproductibilité ?

Reproductibilité

Versionner Partager Archiver

git GitLab zenodo Software Heritage

Créer un environnement virtuel

docker CONDA

Installer des outils par un gestionnaire

CONDA Mamba

Créer un script d'analyses

snakemake nextflow

Déporter l'analyse sur un serveur

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Portabilité de l'exploration des résultats

Shiny quarto rmarkdown jupyter

Éditer un rapport d'analyses

quarto rmarkdown jupyter

Merci pour votre attention

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

cnrs INRAE Inserm cea INVESTIR L'AVENIR